Interactions Between the Gene Ontology and a Domain Corpus for a Biological Natural Language Processing Application

Cornelia M. Verspoor verspoor@lanl.gov
Cliff Joslyn joslyn@lanl.gov
George Papcun gjp@lanl.gov

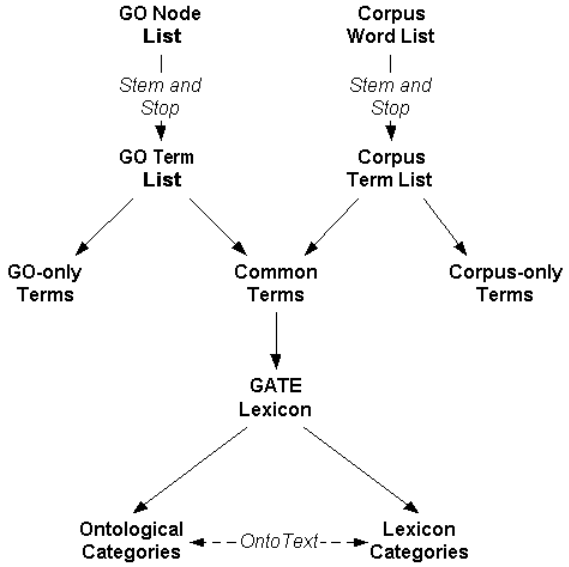
Los Alamos National Laboratory
Computer & Computational Science Division
PO Box 1663, MS B256
Los Alamos, NM 87545

1 Introduction

In any natural language processing (NLP) application, there is a critical need to manage lexical resources in a manner which supports representation of syntactic and semantic constraints on lexical use. In domains which contain much highly specific terminology, such as the biological domain, it is often a daunting task to construct such lexical resources. We turn, therefore, to existing terminological and ontological resources for the domain. However, while there is significant overlap in the requirements for an NLP system with those of ontological data representations, the requirements are not identical. It is important to consider with care the integration of a lexicon with an ontology into a single application.

Specifically, an NLP system is heavily focused on terminological management issues. Words which are synonymous from the perspective of a given ontology may behave quite differently from a linguistic perspective. The internal structure of a multi-word term is largely irrelevant for ontological use, but may be critical in linguistic processing to support recognition of the term in text, where there may be intervening words or surface variations not captured in the ontology. However, the semantic grounding provided by an ontology can be extremely important for enabling precise analysis of the meaning conveyed in relevant text sources.

We discuss a prototype system, currently under development, that aims to extract regulatory relationships from biological text (Papcun et al 2003), and which depends on the existence of domain-specific lexical resources. While our customer has supplied some lists of terms that are associated with particular semantic types, these lists are invariably incomplete and exist independently of any domain ontology. We therefore turn to the Gene Ontology (GO, http://www.geneontology.org) (Ashburner et al 2000) as a source of richer semantic data for lexical resources. The architecture we follow for construction of those resources is shown in the figure below. Term lists are derived from the GO and a customer-supplied public text corpus respectively, and then stemmed in order to determine distinct term lists. We maintain multi-word terms as phrases in addition to breaking them down into terms consisting of individual words. Finally, certain terms considered to be uninteresting (stop words), including linguistic function words and extremely frequent words, are eliminated from the lists. Terms held in common to GO and the corpus are extracted as the lexicon for our system. The result is a lexicon in which terms can be directly associated with the semantic categories of the domain ontology.



2 GO as a source of lexical data

As a controlled vocabulary, the GO provides an important source of domain-specific terminology that can be used to inform lexicon development for an NLP system. It can be used in the following ways:

- Ontological relations represented in the GO can be reasoned upon in combination with linguistic analysis in order to establish ontological relations among individual terms. We see an example of this type of processing in the second figure, in which relations between heads of phrases are inferred from the relation between the phrases as a whole, e.g. that lipidation is a kind of biosynthesis. We are exploring the extent to which relations in the GO can be exploited

in establishing relations between individual terms in the lexicon.

- The hierarchical structure of the GO can be exploited to represent semantic constraints and generalizations in linguistic rules, since each term derived from the GO is associated with a node in the ontology. For instance, a rule may require that a particular argument be some type of protein metabolism. With reference to the GO, we can verify that this holds for a given phrase identified in the text. These types of constraints allow us to more accurately identify particular relationships.
- Definitions of terms in the GO can be used to establish additional lexical relations; words which are used to define a given word can be assumed to have a contextual relationship with that word. This in turn can be used in the NLP system to support word sense disambiguation in the face of words with multiple meanings or in the case of overlapping multi-word units. This is in the spirit of word sense disambiguation work based on machine readable dictionaries (Lesk, 1986).
- Multi-word phrases occurring as nodes in the GO may correspond to non-decomposable word sequences that can be recognized during linguistic parsing to improve structural analysis.

3 Text as a source of ontological data

The corpus of domain texts can also be viewed as a source of ontological data that may or may not be represented in the reference ontology. To the extent that the corpus contains information not captured by the ontology, the ontology may be insufficient (depending on its intended purpose). We are exploring the use of NLP technologies to identify ontological relations expressed in the corpus. These relations would be proposed for integration with the ontology, such that it becomes congruent with the corpus. The implemented techniques would draw on the lexicon, so this represents a feedback loop between the ontology and the NLP system.

4 Integration into the NLP system

The lexicon resulting from intersecting the GO with the domain corpus is represented in terms of gazetteers (term lists) in the General Architecture for Text Engineering (GATE) framework (http://gate.ac.uk). GATE itself only supports the assignment of major and minor types to a given list of lexical items, as shown in the second figure. This alone does not provide sufficient semantic granularity to enable precise relation extraction, and furthermore does not allow us to take advantage of the semantic structure provided by the grounding of the terms in the GO. We therefore incorporate extensions to GATE provided by OntoText Lab (http://www.ontotext.com) which allow us to define mappings of ontological categories from GO to lexical features in the GATE lexicon. With this in place, lexical items can be considered by the NLP system in the far richer semantic context provided by the GO.

5 Acknowledgements

This work was funded in part through a Los Alamos National Laboratory collaboration with Procter & Gamble Corporation.

6 References

Ashburner, M; Ball, C.A.; and Blake, J.A. et al (2000). "Gene Ontology: Tool for the Unification of Biology", *Nature Genetics*, v. 25:1, pp 25-29.

Lesk, M.E. (1986). Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the Fifth International Conference on Systems Documentation*, Toronto, CA, pp. 24-26, ACM.

Papcun, George, Kari Sentz, Andy Fulmer, Jun Xu, Olaf Lubeck, and Murray Wolinsky (2003). A construction grammar approach to extracting regulatory relationships from biological literature. *Pacific Symposium on Biocomputing 2003*, Kauai, Hawaii.